Combining simple interactivity and machine learning: a separable deep learning approach to subthalamic nucleus localization and segmentation in MRI for deep brain stimulation surgical planning

John S H Baxter, Pierre Jannin

Combining simple interactivity and machine learning: a separable deep learning approach to subthalamic nucleus localization and segmentation in MRI for deep brain stimulation surgical planning

John S. H. Baxter* **and Pierre Jannin**
Université de Rennes 1, Laboratoire Traitement du Signal et de l'Image (INSERM UMR 1099), Rennes, France

Abstract

Purpose: Deep brain stimulation (DBS) is an interventional treatment for some neurological and neurodegenerative diseases. For example, in Parkinson's disease, DBS electrodes are positioned at particular locations within the basal ganglia to alleviate the patient's motor symptoms. These interventions depend greatly on a preoperative planning stage in which potential targets and electrode trajectories are identified in a preoperative MRI. Due to the small size and low contrast of targets such as the subthalamic nucleus (STN), their segmentation is a difficult task. Machine learning provides a potential avenue for development, but it has difficulty in segmenting such small structures in volumetric images due to additional problems such as segmentation class imbalance.

Approach: We present a two-stage separable learning workflow for STN segmentation consisting of a localization step that detects the STN and crops the image to a small region and a segmentation step that delineates the structure within that region. The goal of this decoupling is to improve accuracy and efficiency and to provide an intermediate representation that can be easily corrected by a clinical user. This correction capability was then studied through a human–computer interaction experiment with seven novice participants and one expert neurosurgeon.

Results: Our two-step segmentation significantly outperforms the comparative registration-based method currently used in clinic and approaches the fundamental limit on variability due to the image resolution. In addition, the human–computer interaction experiment shows that the additional interaction mechanism allowed by separating STN segmentation into two steps significantly improves the users' ability to correct errors and further improves performance.

Conclusions: Our method shows that separable learning not only is feasible for fully automatic STN segmentation but also leads to improved interactivity that can ease its translation into clinical use.

 [DOI: 10.1117/1.JMI.9.4.045001]

Keywords: subthalamic nucleus localization; subthalamic nucleus segmentation; convolutional neural networks; deep brain stimulation; separable machine learning; human–computer interaction.

Paper 21311GRR received Nov. 30, 2021; accepted for publication Jun. 16, 2022; published online Jul. 11, 2022.

1 Introduction

Deep brain stimulation (DBS) is a common method for the treatment of Parkinson's disease (PD) as well as an increasingly used method for addressing the symptoms of other neurological disorders such as epilepsy. The critical aspect of DBS is the accurate positioning of stimulation electrodes at a particular anatomy of interest determined in a preoperative planning stage that

*Address all correspondence to John S. H. Baxter, jbaxter@univ-rennes1.fr

2329-4302/2022/$28.00

combines imaging and symptomatology information.[1] Due to the complexity of determining the appropriate electrode trajectory, the anatomies of interest must be accurately segmented from preoperative images.

For DBS preoperative planning, segmentation is predominantly determined via registration of the patient images into an atlas-space in which the anatomy of interest, often the subthalamic nuclei (STN), as well as other salient areas, have been presegmented.[2,3] The use of a presegmented atlas has several advantages. From a clinical point of view, it is possible to port a large number of segmented regions from the atlas into the patient-space, simplifying the computational aspect of the workflow. From a research perspective, using an atlas, patient-specific information from the patient images can be ported back into a common atlas co-ordinate system, allowing for populationwise information to be discerned, which could help guide treatment.[4] The main issue with this form of segmentation, however, is that the registration process must be deformable but sensitive to the relatively low-contrast and small subcortical gray matter structures despite their proximity to the larger and more salient ventricles. In addition, deformable registration is computationally intensive and sensitive to parameterization.

Thus, a deep learning approach that segments the left and right STNs would be highly beneficial from a workflow point of view, rendering the segmentation more efficient and robust to changes in the underlying image. Convolutional neural networks (CNNs) have become an increasingly popular tool for many aspects of cranial MRI via deep learning, having a number of desirable properties (such as translation invariance for its lowermost layers) and because of the large variety of network structures that make use of them.[5] However, directly segmenting the STN from the full volumetric images is problematic as they are small with low salience compared with "distractors," which represent anatomy that is visually complex with higher contrast (and thus may require more computational power to detect) but does not provide direct information about the STN segmentation itself. (For example, cortical gray/white matter interfaces could be visually similar to the boundary of the STN, but they do not provide any information about it.) In addition, it is necessary to use relatively high-resolution images for segmentation, requiring a large amount of memory for even basic CNNs, with the vast majority (well over 99% of voxels) not being inside or even proximal to the STN.

The use of CNNs themselves is not new to STN segmentation. Previous work by Milletari et al.[6] investigated several different CNN architectures on this particular problem, using patch voting to localize the STN. The networks were designed to be sufficiently lightweight that the convolutional operators could be applied to whole images without overwhelming their computational resources. Although their two-dimensional (2D) methods experienced average accuracies of under 20% Dice, their volumetric method substantially surpassed that to an average of 61.4% Dice for correctly localized STNs. The downside of such a patch-based framework is that the patch-voting procedure may spuriously identify the STN in other regions of the brain, an event that Milletari et al. called a failure. In their method, a failure rate of between 5% and 10% in terms of all subcortical structures was achieved. However, Milletari et al. did not separate this result per structure, and the set of segmented structures included several larger and more salient structures where one would intuitively expect a lower probability of failure, which suggests that the failure rate may be even higher for the STN in particular.

These approaches tend to use a more traditional end-to-end learning approach in which a singular learnable component (i.e., neural network) is constructed for a singular task, with other tasks being largely accomplished by pre- or postprocessing. The advantage of this is largely its simplicity, with any coupling between learned concepts being handled automatically within the singular learned network. For example, in the method proposed by Milletari et al.,[6] the network must learn representations that encode not only information about whether or not an individual voxel is within a structure but also more global information regarding the direction from the voxel to the center of the structure, and it is left to the algorithm to determine how much computational resources are dedicated to each task. This does have fundamental disadvantages though as it often means larger amounts of data are necessary (the cited previous approaches use databases consisting of one hundred or more annotated patient images) for the network to understand this coupling unless the network itself (along with pre- and postprocessing) is designed in some manner to assist with this distribution, conceptually moving toward separable learning. The fundamental main difference between separable learning and this is the separation

of training these structures, which itself is beneficial as training could be distributed more readily across multiple resources and updated independently. In addition, this leads to less wasted training time if the concepts fundamentally rely on each other, that is, if the first learning component fails, then the second must as well. Until sufficient training has occurred so thatthe first component does not provide any meaningful, nonrandom predictions, the second component effectively would experience a "garbage-in-garbage-out" scenario. Thus, separable learning saves on this otherwise wasted effort by approximating the final prediction quality of earlier learned components, rather than waiting for these components to reach this accuracy and then training the downstream components separately using these approximations.

Leveraging 7 Tesla B_0 field strength (i.e., 7T) images toward the more clinically available 3T ones was explored by Milchenko et al.[7] and Kim et al.[8] who used a collection of seven presegmented 7T images as a multi-atlas. By having multiple users segment the 7T images, they were also able to determine approximate inter-rater variabilities for STN segmentation (although at a higher resolution and contrast than at 3T) of between 55% and 71% Dice coefficient, indicating the difficulty of this problem and providing a rough benchmark for determining STN segmentation quality. This variability was highlighted by Duchin et al.[9] on 7T images showing a high patient variability in terms of the STN's volume and extent. Both Milchenko et al. and Kim et al. recognized the difficulty of directly segmenting 3T clinical images, hence the reliance on higher contrast clinical atlases, which are problematic themselves due to high-field image distortion. Although this distortion is more pronounced in cortical regions, Lau et al.[10] found that the geometric distortion between 3T and 7T MRI is on the order of 1 to 2 mm around the subthalamic nucleus, rendering it more difficult to directly use 7T image registration given the lack of contrast in the 3T image to correct for these local distortions. Zhao et al.[11] combined the two approaches and applied U-Nets to directly segmenting the STN on high-resolution susceptibility maps and achieved an accuracy on the order of 78% Dice, although they only used healthy subjects.

1.1 *Contributions*

The goal of this paper is to develop an efficient deep learning approach to STN segmentation using traditional CNN components. To overcome the issue of maintain resolution while limiting the required computational memory required for training the network, the segmentation is split into two separate pieces, each with a distinct focus. The first is a lightweight network designed to work with the full volumetric image, but only to roughly estimate the location of the left and right STNs. The second network then uses heavily cropped images centered at that location, segmenting the STN located within. As the cropped images are orders of magnitude smaller than the full volumetric image, the problem of inherent class imbalance is highly mitigated and a larger amount of computational power can be dedicated to it without overwhelming the computational memory. To the best of the author's knowledge, this is the first separable machine learning infrastructure designed to segment specifically the STN in clinical quality MR images and thus the first to investigate human–computer interaction (HCI) aspects in this particular context.

2 Materials and Methods

2.1 *Images*

Ten patients ($F = 4$, $M = 6$) were extracted from the pyDBS database[2] with corresponding T1-weighted (T1w) (Phillips Achieva 3T, GRE with TR = 11, TE = 4.6, FA = 15) and T2-weighted (T2w) (Phillips Achieva 3T, T2 TSE with TR = 3035, TE = 80, ETL = 15, 2 averages) images both with isotropic 1 mm spacing. The collection and use of this data was approved by the Institutional Research Ethics Board. The T2w images were resampled into their corresponding T1w image space, with the T1w image space being consistently larger than that of the T2w image. They were then stored in RAS orientation and symmetrically zero-padded to a size of $256 \times 256 \times 192$ voxels to maintain a consistent image size. The images, having been collected at the same time and resampled into the same space, are considered to be coregistered, and no registration errors were observed by the clinical expert.

2.2 *Registration Approach*

To contextualize the performance of the proposed framework, a comparative method is introduced. The current clinical state-of-the-art method in DBS preoperative planning is the use of deformable registration, porting the segmentation from a preannotated image to the current patient image.[2] The comparative approach uses the ParkMedAtlis atlas version 3,[4] which is registered to the patient image using the Advanced Normalization Toolkit (ANTs).[12]

The first step is skull-stripping the image using BrainVisa[13] to acquire the brain mask. The initial rigid and then affine registration are acquired using mutual information within the brain mask. This allows for an estimated subcortical region mask to be derived. This mask is then used to update the affine registration using mutal information within the subcortical mask. The final deformable portion used ANTs SyN[12] using cross-correlation as the metric. The use of a subcortical mask to refine the affine registration is required to ensure that the registration has adequate performance for subcortical structures, rather than aligning the more salient, but more variable, cortical ones and thus providing a better initialization for the deformable registration step in the region surrounding the STN and other structures of interest.

2.3 *Ground Truth Segmentation*

To determine ground truth segmentations, a clinical expert modified the atlas-based segmentations to agree to salient edges in the T2w MRI using ITKSnap in which both the T1w and T2w images could be readily visualized in tandem. The use of the atlas-based segmentations as an initial guess at the ground truth allowed for the variability between manual segmentations to be more controlled. For each patient, the T1-weighted image was deformably registered to the version 3 of the ParkMedAtlis atlas[4] as described in Sec. 2.2.

2.4 *STN Localization*

The network architecture is shown in Fig. 1. This architecture is designed to be multiresolution, that is, to contain a series of levels that process the image information at a particular pixel size. This is conceptually similar to networks such as U-Nets[15] in that different levels address the machine learning problem using information present at different granularities.

The crucial aspect of this network that makes it usable is the sampling operator that, at each level, restricts the attention of the network to a smaller area in which it believes the STN is located. This massively reduces the amount of memory needed as the region sampled is orders of magnitude smaller in volume that the original image at the finest-resolved levels of the network. However, this operator is highly nondifferentiable with respect to the co-ordinates of the center of the region, meaning that the network cannot use the finest-resolution estimate as the final estimate in training as the gradients would not be able to propagate to coarser-resolution layers. To overcome this difficulty, the estimates are assigned a non-negative weight (that sum to 1) for each resolution level, allowing the error gradients to immediately flow to each resolution level simultaneously.

For training, data augmentation including in-plane rotations (std. 10 deg) and three-dimensional translations (std. 8.33 mm iso. or 5 mm in each direction) was applied to each dataset 50 times in each epoch. The localization network was trained from scratch with a batch size of 16. The Adam optimizer was used with a learning rate of 10^{-3} and an exponential decay rate of 0.05 per epoch.

The training was performed in a leave-one-out cross-validation style in which one dataset was used for testing, one of the remaining datasets was randomly selected as the validation dataset (to determine the number of training epochs), and the remainder were used for training the network weights. The network with the best performance on the validation dataset was then applied to the testing dataset and saved for further experiments.

2.5 *STN Segmentation*

Given the approximate locations from the previous network, the image can be cropped to two single patches, ideally centered on each of the left and right STNs. The patch size chosen was

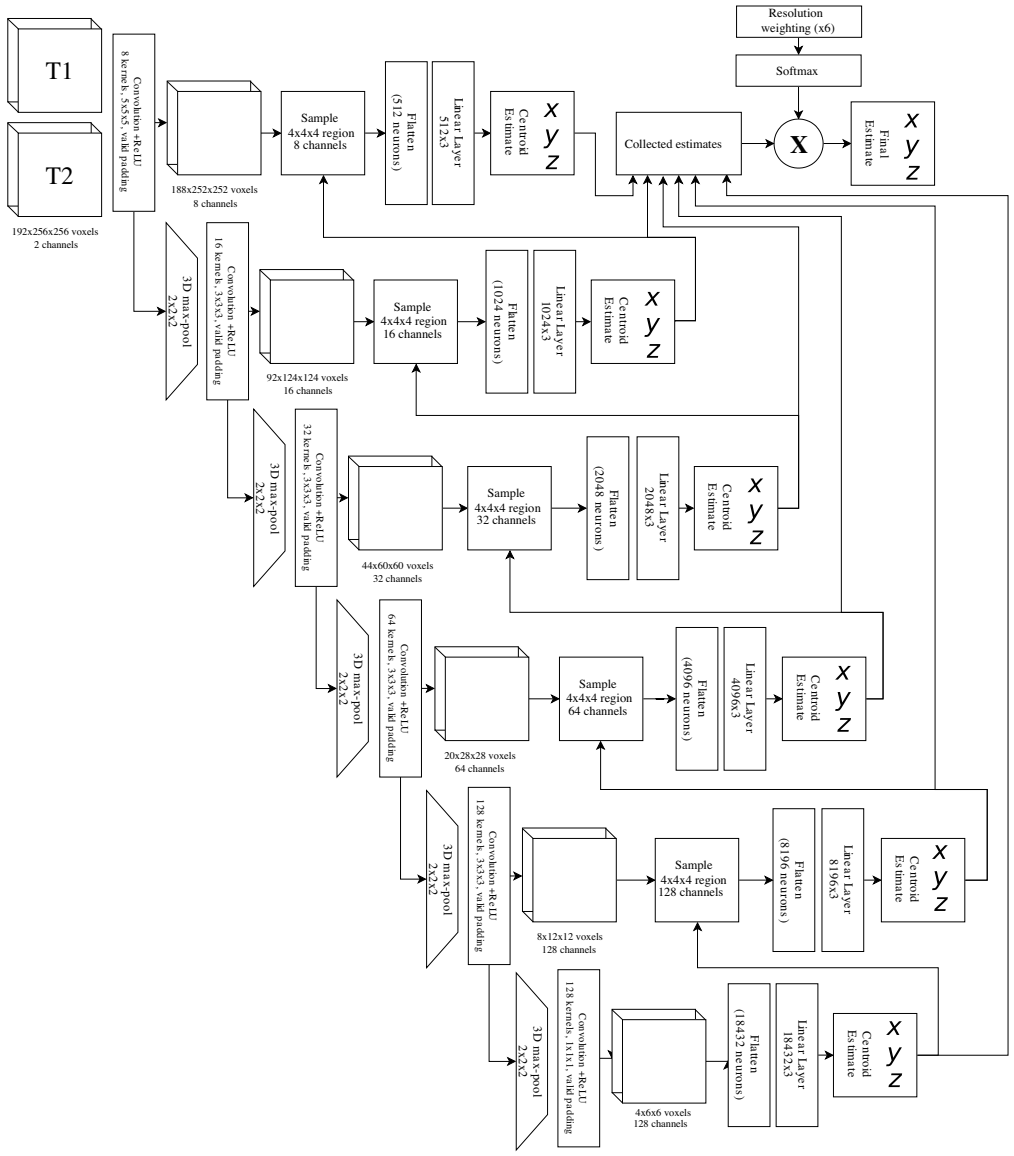

Fig. 1 Multiresolution network for the localization of the STN (replicated from Ref. 14).

$24 \times 24 \times 24$ voxels, which allows for $\sim 3\sigma$ error to occur in either direction while maintaining the entirety of the STN within the volume. As with the previous system, dataset augmentation (translation std. 3 mm in each dimension, rotation std. 10 deg) is used to increase the network's robustness both to the input image and to localization errors.

Because of the greatly reduced image size, a more traditional V-Net[16] style architecture could be used without overwhelming the memory capabilities of the learning system. There are three principle differences between our network and the original V-Net:

1. The dense feature stacks are implemented not by concatenation but by maintaining a larger number of independent convolutions that are added in-place followed by appending the output to the list of inputs. This minimizes the amount of memory required by minimizing the number of concatenation operators that duplicate the information in the tensors. The difference between this memory-efficient dense layer and a regular dense layer is shown in Fig. 2.
2. The pooling operators, instead of using strided convolution, use a "consolidation" convolution followed by max-pooling. This consolidation convolution is similar to the memory-efficient dense layer, taking a list of inputs that have a kernel applied to each

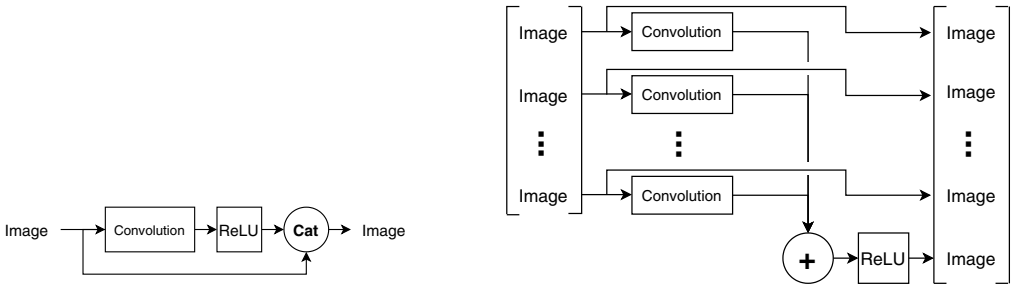

Fig. 2 Difference between regular dense layer implemented using tensor concatenation and a single convolution filter and an equivalent version that uses lists of convolution kernels, in-place addition, and finally list concatenation, which altogether are more memory efficient.

and then summimg, but instead of appending the result to the list, only the single resultant image is returned. This simplifies the computational graph by ensuring that, before each max-pool, the feature stack is transformed from a list of images into a single image. Together, these act similarly to a strided convolution but with additional nonlinearity.

3. The contribution to the segmentation is directly computed at each level, rather than requiring a final convolution operator to combine them. This again avoids the use of concatenation, allowing for deeper networks.

The goal of these modifications is to allow for more dense layers to be added to the network without quickly exhausting the memory supply on the GPU. The overall architecture is shown in Fig. 3. The loss function used in training is the unweighted combination of the Dice coefficient (both foreground and background) and the mean binary cross-entropy. The binary cross-entropy term allows the network to quickly converge to an approximately correct STN, whereas the two Dice terms allow it to perform fine-tuning.

The segmentation network was also trained from scratch with a batch size of 32 for 100 epochs with 50 augmented versions of each image per epoch. The Adam optimizer was used with a learning rate of 10^{-3}. Again, the training was performed in a leave-one-out cross-validation style in which one dataset was used for testing and the remainder were used for training the network weights. To determine the accuracy of the segmentation networks, the localization from the corresponding localization network (i.e., the one with the same testing

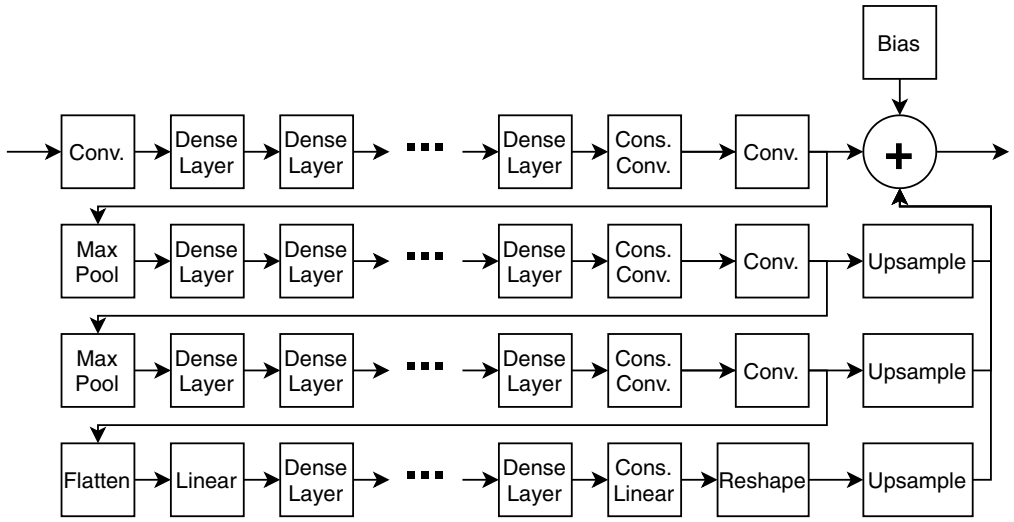

Fig. 3 Segmentation network composed of several resolution levels composed of a series of dense layers [described in Fig. 2(b)] and consolidation convolution (Cons. Conv.) layers. (Replicated from Ref. 17).

dataset) was used to get an accurate representation of the localization error without introducing data leakage.

2.6 *Human–Computer Interaction Study*

One of the primary motivations behind separable learning is to improve the interactive capabilities of machine learning while still being fully automatic. For this framework, in particular, having a separation between the localization and segmentation components of STN segmentation allows the user two opportunities to correct the algorithm. The first is to relocalize the STN, that is, to replace the centroid that is automatically estimated by the network with their own. This means that, in the case of extreme error on the part of the localization network, the user can readily correct it by simply clicking on the STN in the image. The second interaction mechanism is the most common: direct segmentation editing, that is, if the segmentation is largely correct, the user can relabel individual voxels. Often, this is the only interaction mechanism provided to clinicians, meaning that, in the case of very large segmentation errors, the correction process is, in essence, manual segmentation.

To analyze this interaction mechanisms, we created an HCI experiment. In this experiment, the participant is given an automatically generated segmentation of the left and right STNs using the described method although the localization component is given a random isotropic Gaussian-distributed error with a standard deviation of 8.66 mm before applying the segmentation component. This allows the initial automatic segmentations to have Dice coefficients between 0% and 85%. The participant can then correct the segmentation either by changing the localization, i.e., the centroid location (referred to as interaction mechanism A), or by painting over the segmentation, changing the label or voxels one at a time (referred to as interaction mechanism B). Which mechanism is provided is told to the user via text before the interface is launched as well as being encoded in the color of the interface (light red for interaction mechanism A and light blue for interaction mechanism B), and the mechanism is randomly assigned using counterbalancing to control for fatigue. Another interface (light green, shown in Fig. 4) was implemented without editing capabilities. Interaction mechanisms A and B refer to different fundamental types of interaction and thus could likely be seen as complementary in practice.[18]

Seven novice participants (M = 4, F = 3) were recruited and performed eight trials per day for 4 to 10 days. Each of the eight trials used a different image. The ninth image and gold

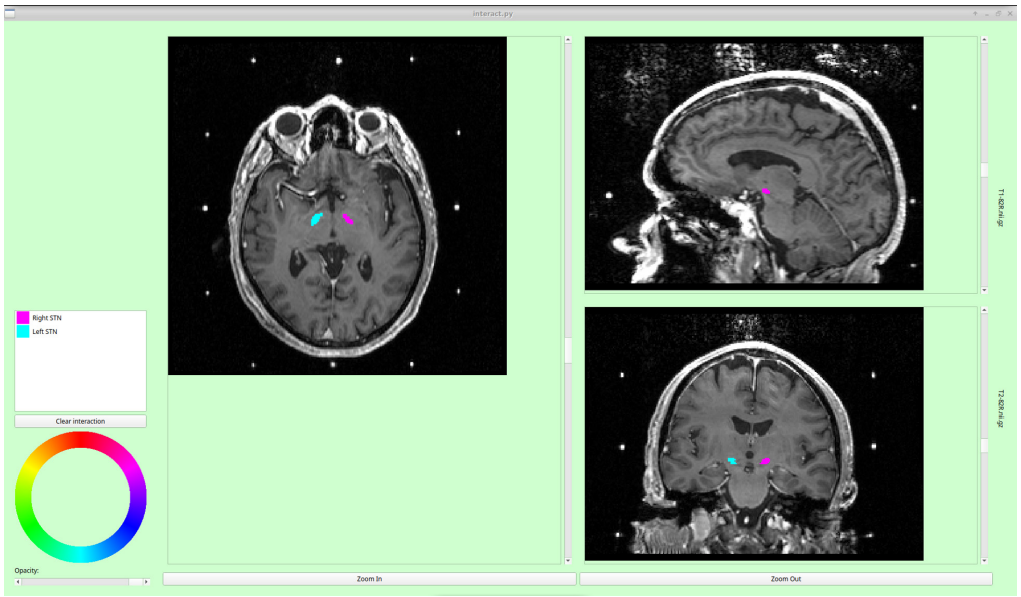

Fig. 4 Interface used in the HCI study. The interface comes in three colors: light green indicating no segmentation editing mechanism (used for habituating the participant to the task), light red indicating relocalization as the sole editing mechanism, and light blue indicating painting as the sole editing mechanism. The left and right STNs are shown in cyan and magenta, respectively.

standard was shown to the participants at the beginning in the light green interface for the participants to acclimate to navigating the interface (e.g., changing slides in the three 2D views, zooming, and switching between T1w and T2w images) and to better understand the anatomy. While the participant interacts with an interface, all of the actions are saved in a log file along with their timestamps and Dice coefficients for the current segmentation. Immediately after interacting with the interface, the participant is presented with an electronic NASA Task Load Index (TLX)[19] form, which is a common tool for quantifying subjective aspects of usability, and these results are also logged.

Our concrete hypotheses are as follows:

1. Interaction mechanism A will result in higher postcorrection Dice coefficients than interaction mechanism B when the initial segmentation Dice coefficients are low, that is, below 40%.
2. Interaction mechanism A will be interpreted as being more usable generally,
3. Response times using interaction mechanism A will be faster than those using interaction mechanism B when correction is required and will be equivalent when correction is not required.

3 Results

3.1 *STN Localization*

Quantitative results for STN localization can be found in Table 1. Overall, the method came within a few millimeteres of the centroid location, indicating that a sufficiently high level of accuracy could be achieved for the purpose of restricting the image to a smaller region of interest surrounding the STN. This method was also found to have statistically significantly better performance than the use of deformable registration, along with being much faster to compute. This improvement in performance verifies our hypothesis that a multiresolution CNN is well-suited for this particular problem, allowing for coarser-resolution levels to roughly localize the STN based on larger surrounding anatomy whereas finer-resolution levels can focus on the particular intensity distribution of the STN, undistracted by the more salient proximal structures.

3.2 *STN Segmentation*

Qualitative results of the segmentation are shown in Fig. 5. Figure 6 shows the improvement in segmentation accuracy resulting from changing from the deformable registration-based result to the neural network constructed at various depths. The Dice improvement is shown rather than the Dice coefficients directly to control for the variability introduced by the difficulty of the particular patient dataset, thus only showing the variability that results from the neural network itself. This also allows for paired Student's *t*-tests to be performed across the two approaches,

Table 1 Quantitative results for estimating the STN centroid location. The Δ column indicates the improvement (i.e., difference) of the proposed multiresolution CNN method over the traditional deformable registration method.

	Multiresolution CNN error (mm)	Deformation registration error (mm)	Δ (mm)	*p*
RL	0.80 ± 0.64	1.21 ± 0.90	0.41 ± 0.86	—
AP	0.94 ± 0.66	1.99 ± 1.43	0.97 ± 0.32	—
SI	1.18 ± 1.02	1.87 ± 1.66	0.69 ± 0.64	—
Total	2.03 ± 0.81	3.39 ± 1.70	1.36 ± 0.63	0.29%

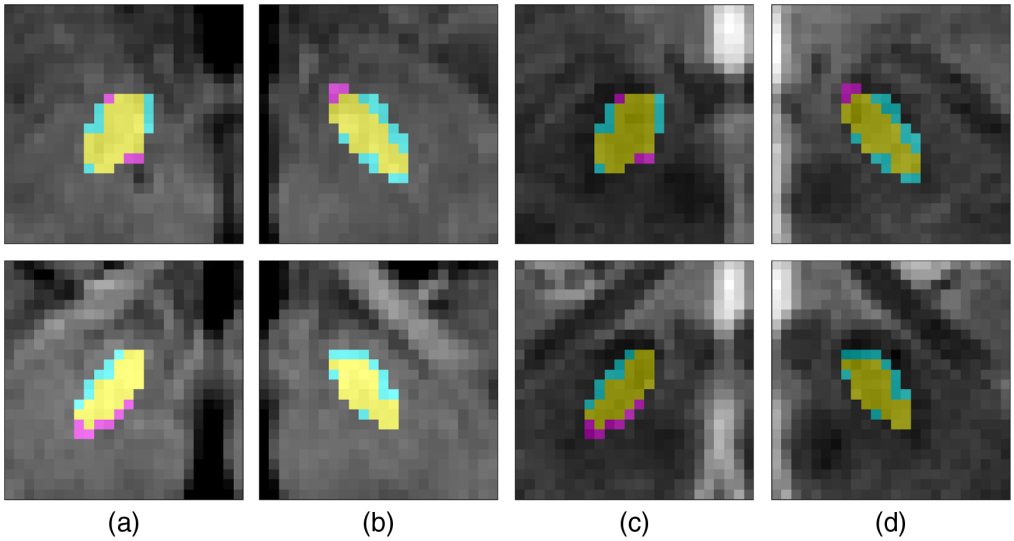

Fig. 5 Qualitative results are shown for two patients. The T1w MR images cropped to the STN region displaying the proposed method overlapped with the manual segmentation are shown in columns (a) and (b) for the left and right STNs, respectively. The same for the T2w images is shown in columns (c) and (d) for the left and right STNs, respectively. For each image, the manual segmentation is shown in magenta, the automatic segmentation in cyan, and the overlap in yellow.

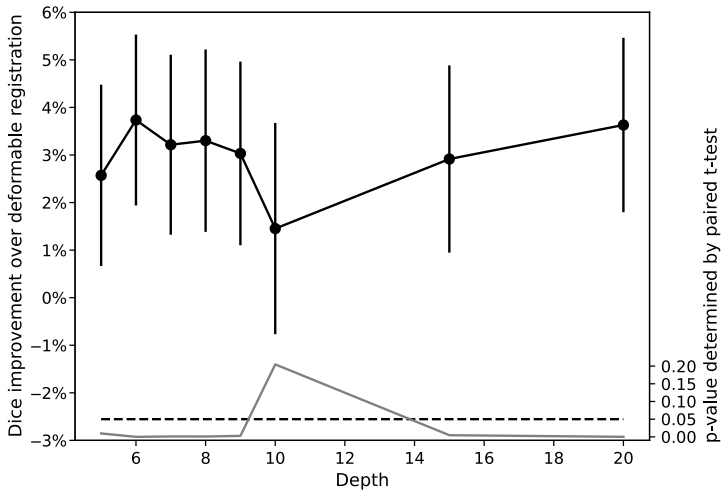

Fig. 6 Improvement in Dice coefficients comparing the network with deformable registration (solid black line, *y*-axis to the left). The curve shows a distinct double descent with an early peak at a depth of 6, the worst performance at a depth of 10, but improved performance after that depth. *p*-values determined by the student's *t*-test are shown below (solid gray line, *y*-axis to the right) with a dashed line marking the 5% level.

noting that, with the exception of depth 10, all results are statistically significant after Holm–Bonferroni correction.

To indicate an acceptable level of error, we can compare them against the Dice coefficients generated by offsetting the manual segmentations by a small amount, specifically a 1 voxel (1 mm) dilation and a 1 voxel shift in any of the cardinal directions to simulate potential variabilities such as subvoxel coregistration error between the T1w and T2w images. (An example of this is shown in Fig. 7). These errors are likely higher than the expected interoperator variability but indicate what Dice coefficients are globally subvoxel and can be computed directly from the reference segmentations available in our dataset. The results of this evaluation are given in Table 2. The proposed neural network significantly outperformed each of these offset

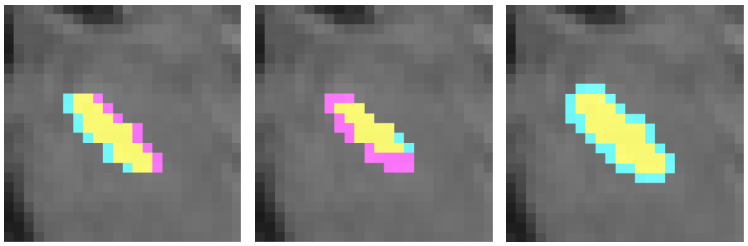

Fig. 7 Example perturbation for shifting in the left (a) and superior (b) directions and for dilation. For each, the manual segmentation is shown in magenta, the perturbation in cyan, and the overlap in yellow.

Table 2 Quantitative results for segmenting the STN.

($n = 20$)	Dice coefficient	p-value versus Prop.	p-value versus Reg.
Proposed CNN (depth 6)	61.9 ± 9.5%	—	—
Deformable registration-based	56.3 ± 9.7%	$\mathbf{5.3 \times 10^{-4}}$	—
1 mm dilation	50.5 ± 2.8%	$\mathbf{2.1 \times 10^{-32}}$	$\mathbf{2.8 \times 10^{-4}}$
1 mm LR shift	54.2 ± 8.1%	$\mathbf{5.4 \times 10^{-6}}$	8.1×10^{-2}
1 mm AP shift	53.0 ± 6.8%	$\mathbf{5.3 \times 10^{-8}}$	1.3×10^{-2}
1 mm SI shift	55.0 ± 5.5%	$\mathbf{2.0 \times 10^{-7}}$	7.7×10^{-2}

The p-values are determined via a two-factor ANOVA test in which the STN (patient as well as side) is one factor and the other is the type (i.e., proposed method, registration-based, etc.). The p-value given is for the difference between the methods and statistically significant results (after Holm–Bonferroni correction with a threshold of 5%) are shown in bold.

segmentations, indicating that it is within a general accuracy of 1 voxel, which likely renders it within the variability expected in the manual segmentation due to partial volume effects and small patient motion. The registration approach was significantly better for some, but not all, of these offsets, indicating that its performance is comparable to having a 1 voxel accuracy but is less likely to be within the variability of the manual segmentation itself. Due to the very small size of the STN, it is necessary to contextualize Dice accuracy in terms of the possible variability of the reference segmentations. Previous results in the literature suggested that interoperator variability for STN segmentation is 63% Dice, indicating this as the level of approximate human performance for this task.[20]

3.3 *Human–Computer Interaction Study*

Figure 8 shows the relationship between the initial segmentation quality and the quality of the corrected segmentation, both measured in Dice, for both interaction mechanism A (red) and interaction mechanism B (blue). The black line is the 45 deg line, showing when the initial and final segmentations were of the same quality, and the red and blue lines are the regression lines for interaction mechanisms A and B, respectively. Notably, for many of the results in which the initial quality had a high Dice coefficient (i.e., higher than 50%), the responses fall on this line, principally due to the participant determining that the segmentation was "close enough" and exiting the interface (which we call a skip).

From Fig. 8, it appears that interaction mechanism A tended to result in higher performing segmentations than interaction mechanism B regardless of the initial segmentation quality. This general observation was confirmed using a Wilcoxon rank-sum test with details shown in

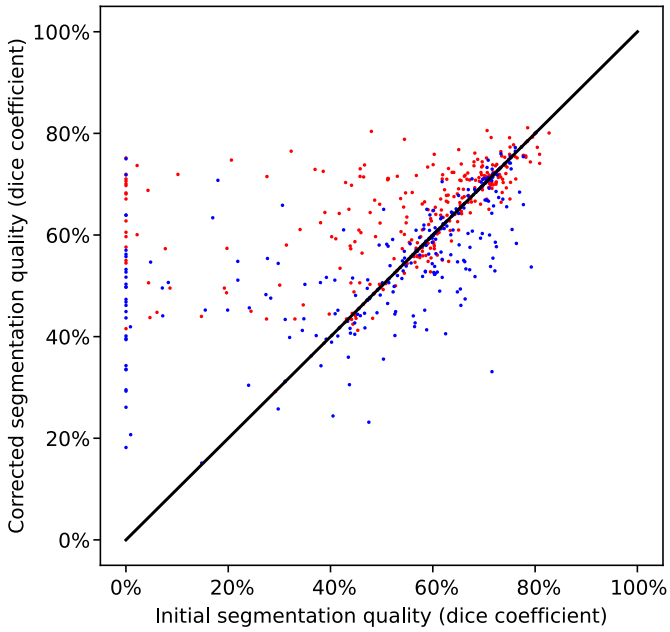

Fig. 8 Scatter plot showing the relationship between the initial segmentation quality and final segmentation quality. Red dots indicate trials performed using mechanism A and blue dots indicate trials performed using mechanism B.

Table 3 Quantitative results for the final segmentation Dice coefficients.

Low initial Dice (<40%)			High initial Dice (≥40%)		
Mechanism A	Mechanism B	*p*-value	Mechanism A	Mechanism B	*p*-value
$58.6 \pm 12.0\%$	$45.4 \pm 12.6\%$	3.3×10^{-6}	$65.1 \pm 9.0\%$	$59.2 \pm 10.5\%$	1.8×10^{-11}

Table 3. Notably, this confirms our original hypothesis, as well as a stronger version that interaction mechanism A also outperformed interaction mechanism B even when the initial segmentation quality was high.

In terms of time, Fig. 9 shows the amount of time taken for each interaction mechanism. This is separated between when the user actively corrected the segmentation versus a skip. The difference between the two methods when correction is applied is fairly clear (Wilcoxon rank-sum test result, $p = 1.3 \times 10^{-24}$) whereas the difference between the mechanisms for the time taken to skip editing is only barely statistically significant prior to statistical correction (Wilcoxon rank-sum test result, $p = 0.039$) and is thus considered insignificant. This verifies both aspects of our second hypothesis, i.e., that interaction mechanism A will lead to faster segmentation editing than interaction mechanism B when such correction is deemed necessary but will be equivalent otherwise. For interaction mechanism A, the proportion skipped was 10%, whereas for interaction mechanism B it was 17%. This may suggest that the easier interaction mechanism could encourage people to correct the segmentation (i.e., raising the bar for acceptable accuracy), but this result is not statistically significant (χ^2 test, $p = 0.059$).

Figure 10 shows the distribution of NASA TLX results for the two interaction mechanisms. Again, the generally improved scores for interaction mechanism A over interaction mechanism B is confirmed via Wilcoxon rank-sum tests shown in Table 4. This means that, across the board, the use of the new interaction mechanism resulting from separable learning is easier than the more traditional manual correction method. The distributions are also clearly different with interaction mechanism A's distributions being skewed toward lower values (i.e., more usable) and interaction mechanism B's distributions skewing toward middle or higher values with much more heterogeneity.

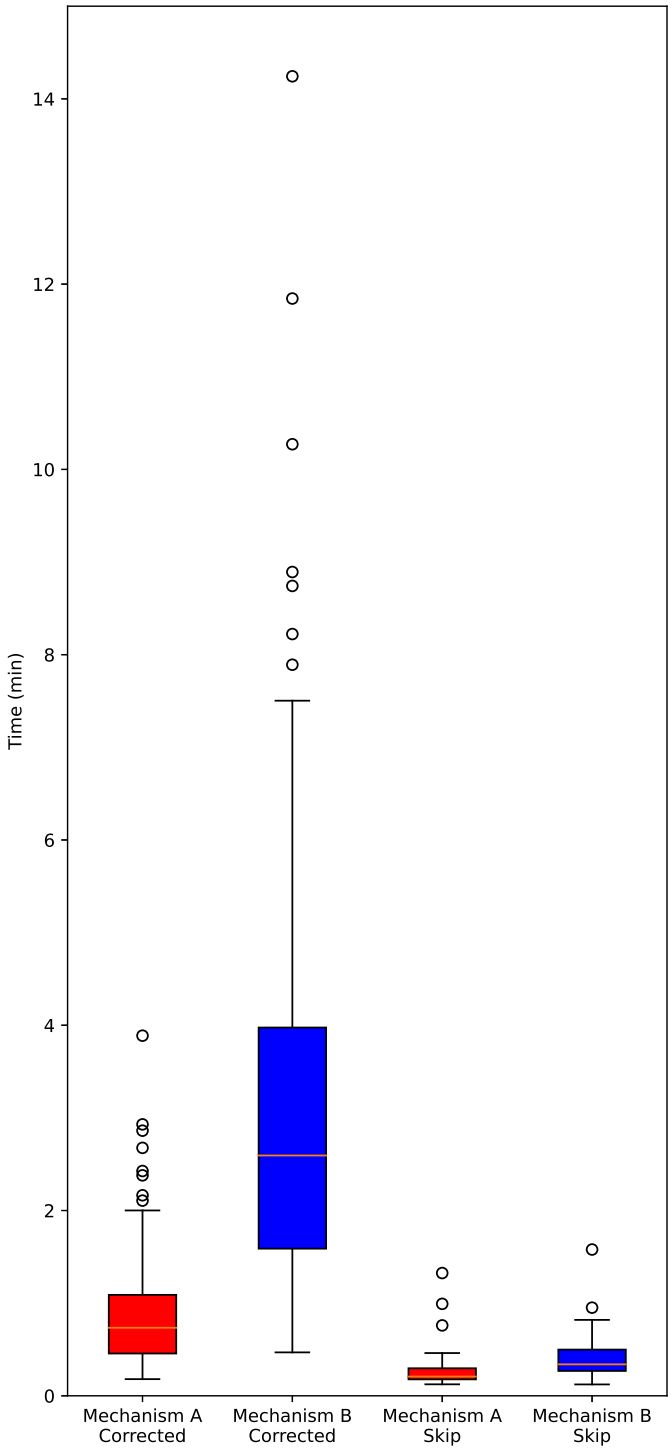

Fig. 9 Time taken in the interface for both trials where the segmentation is corrected and when the correction is skipped.

4 Discussion

4.1 *Segmentation Method*

One interesting aspect of our results is that, for the patch segmentation network, the deep double descent phenomenon[22] can be seen in Fig. 6. This may be largely explained by the structural similarity between the segmentation network proposed and those investigated by Nakkiran

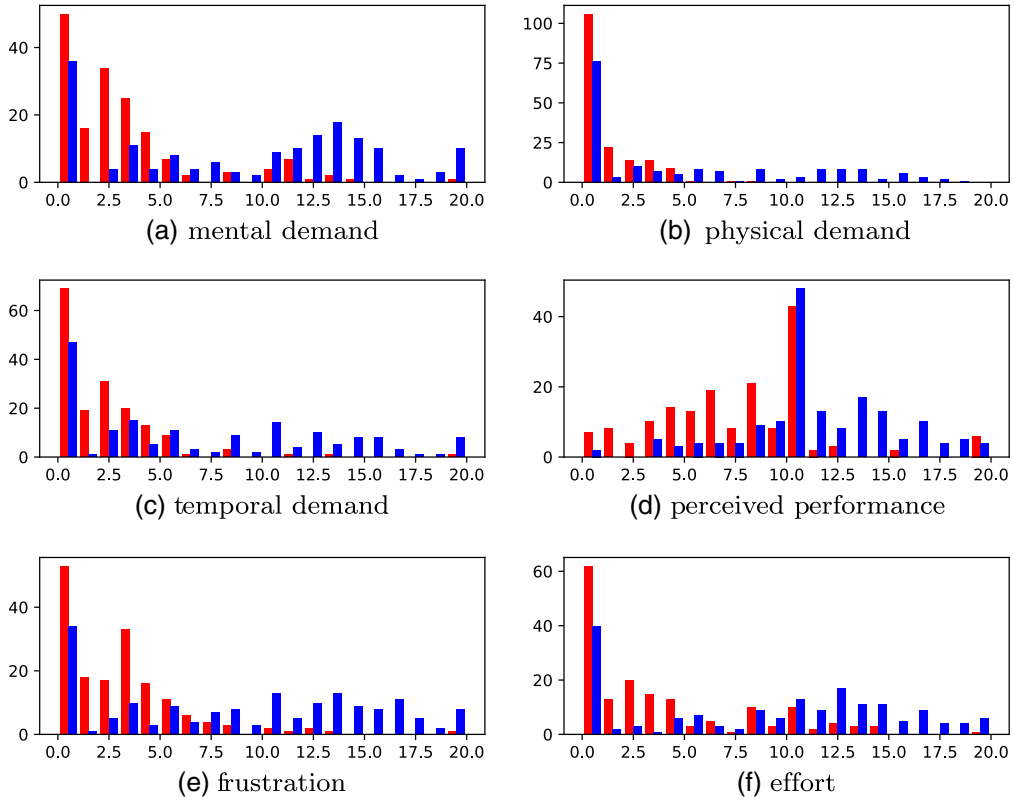

Fig. 10 NASA TLX results for interaction mechanism a (red) and interaction mechanism b (blue). Note that lower values are better for all metrics.

Table 4 Quantitative NASA TLX results. MD, mental demand; PD, physical demand; TD, temporal demand; PP, perceived performance; Fr, frustration; Ef, effort.

	Mechanism A	Mechanism B	p-value
MD	$2.85 \pm 3.46\%$	$8.58 \pm 6.23\%$	2.9×10^{-15}
PD	$0.88 \pm 1.45\%$	$4.59 \pm 5.45\%$	1.5×10^{-7}
TD	$1.86 \pm 2.55\%$	$6.60 \pm 6.11\%$	1.2×10^{-11}
PP	$7.23 \pm 4.06\%$	$11.05 \pm 3.80\%$	9.8×10^{-19}
Fr	$2.65 \pm 3.00\%$	$8.56 \pm 6.22\%$	1.7×10^{-16}
Ef	$3.45 \pm 4.18\%$	$8.54 \pm 6.15\%$	1.8×10^{-12}

et al.:[22] residual-style networks. Although our network uses a dense rather than residual architecture, the underlying characteristic is the same: that the "default state" for an additional layer is centered around the output of the previous layer (i.e., identity) rather than zero, allowing for additional layers to have a mitigated effect on the backpropagation dynamics until they become "useful" for modeling more complex nonlinear behavior. What is particularly interesting about this dataset exhibiting deep double descent behavior is its size, being composed of only 20 datasets (separating the left and right STN) rather than the thousands or millions normally associated with other verifications of the deep double descent hypothesis. Thus, the deep double descent behavior appears to be potentially an inherent property of these types of networks, rather than one conditioned on exceeding a critical mass of data. As far as we can discern, this is the first

observation of the double descent phenomenon in a medical imaging dataset, especially a small one consisting of only 10 images.

One of the primary motivations behind this work was to provide an intermediate representation of the STN segmentation problem that is simultaneously direct and intuitive but does not place a large burden on the clinical user. This fundamentally separates our STN segmentation process into two components, with the coarse localization of the STN being chosen as the intermediate representation. This has the benefit of being (1) easily visualized as a point or glyph could be placed on the image at that location, (2) easily modified by the user who could change it potentially by dragging the glyph or by having a particular key-stroke combination to activate placing it using the mouse, (3) intuitive as it represents the location of the target of interest, and (4) low-burden as the point is not required to be in the STN but in the general vicinity of it. Thus, it represents a much easier mechanism for interacting with the artificial intelligence than requiring the clinical user to manually segment the image or to drag registration control points, which might otherwise be necessary to correct errors in atlas-based segmentation.[18] One important element of future work is to consider and to investigate the human factor components of this interaction mechanism, determining whether or not it is truly more intuitive than possible alternatives and whether clinical users would be amenable to correctable fully automated segmentation methods.

The combination of localization and segmentation is clearly important for STN segmentation. Previous methods have tended to rely either on atlases to provide global context[7,23] or semi-automated methods in which the user-provided initialization encodes the STN's general position.[24] Once localized, our method has a Dice overlap coefficient of ~61.9%, which puts it within the realm of interoperator variability.[20] Given the availability of only a single manual segmentation as a reference standard, it is impossible to gauge if the method has better-than-human performance, although it still would still benefit from increased speed and consistency while still giving the opportunity for correction to the clinical user.

In addition, the manual segmentations were initialized using the comparative deformable atlas-based approach. Although this was done to conserve the time of the expert annotator, it may lead to a slight bias in favor of the atlas-based results.

4.2 *Comparison to Other Fully Automatic Segmentation Algorithms*

Table 5 summarizes recent methods from the literature for fully automatic STN segmentation. As the datasets used in these evaluations are different and the validation techniques used are different, the precise Dice coefficients are not directly comparable; however, it does illustrate that both the proposed method and the comparative atlas-based method fall within the range expected by

Table 5 Dice results for STN segmentation from whole clinical-strength MR images from the literature.

Method	Brief description	No. of patients	Dice (%)
7T registration[21]	Registration with regression forests	10	61 ± 12 (left)
			56 ± 15 (right)
U-Net on high-resolution susceptibility images[11]	Full images at multiple resolutions	75	77.7 ± 6.6
Hough-CNN[6]	Patch-based deep-learning	55	61.9 (left)
			60.9 (right)
Brain-Lab elements[20]	Deformable registration based	30	54 ± 12
DISTAL Atlas[20]	Deformable registration based	30	59 ± 13
Horn electrophysiological atlas[20]	Deformable registration based	30	52 ± 14

the literature. Of the methods in Table 5, the highest performing method, Hough-CNN, is also the other one not to be evaluated against a manual segmentation, but rather a deformable atlas, which may lead to elevated Dice values.

However, it should also be noted that none of the methods listed above lend themselves to being computed in seconds or easily corrected by the user, a core strength of the proposed approach.

4.3 *Human–Computer Interaction*

In terms of time and accuracy, the use of relocalization as an interaction mechanism showed significant improvement over the more traditional correction method, which confirms our motivation regarding the increased interactivity allowed by separable learning. The accuracy is also relatively robust, being much higher than the 1 mm perturbations (Table 1) regardless of the initial segmentation being a high or low quality, indicating that even novice users are perfectly capable of determinining the STN location to within the threshold usable by the segmentation network. In addition, the NASA TLX scores also generally indicated that the relocalization mechanism was easy to use as evidenced by their distribution being heavily skewed toward 0 (i.e., the best value) for all metrics except perceived performance. For that metric, both methods had a mode at 10, the middle of the scale and the default value in the online NASA TLX form, which possibly indicates some uncertainty in the novice population about how well they performed, especially as no quantitative performance feedback was provided to them.

One of the limitations of our HCI study is the use of novices as participants as they are not full experts in STN segmentation. The quality of their segmentations (both manual and automatic) is close to that of expert variability as measured by Polanski et al.[20] (63% Dice) when the initial segmentation is of high quality. With a low Dice, the results of mechanism B do not approach this value, indicating a difference between the novice population and the two experts from Polanski et al.'s study. Our interface was designed in collaboration with an expert neurosurgeon who found the relocalization tool to be extremely useful, although also pointing out the necessity of providing both interaction mechanisms in a clinical interface to use relocalization to correct for large errors and painting to correct for small differences.

The other limitation of our study is that it coupled interaction for both the left and right STNs. This means that we measured accuracy results for each, but timing and NASA TLX results for both were combined. This means that the timing and ease-of-use results could not be split into high initial accuracy and low initial accuracy groups cleanly as the left STN may have been segmented initially with high accuracy and the right STN with low accuracy or vice versa.

4.4 *Placement in Clinical Workflow*

DBS involves a complex workflow with multiple steps notably including the pre- and intraoperative identification/segmentation of the STN for surgical navigation and targeting, respectively.[25] These are conceptually separate problems, relying on different modalities and having different constraints, goals, and requirements.[26]

Our work aims to address the delineation of the STN target at the preoperative stage, which allows for an ideal electrode trajectory to be defined. Due to the small size of the target and the large distance from the skull to the subcortical anatomy, highly accurate segmentation is necessary to ensure that a trajectory that robustly passes through the STN is selected. However, there are additional concerns aside from the overall accuracy of the method, notably the ability to correct errors. Due to the large heterogeneity of the population receiving DBS treatment, there is a high probability that any algorithm will fail for some patient. Thus, having a framework that allows for the automatic segmentation to be quickly and almost effortlessly corrected would be highly desirable for the patients.

Even with high preoperative segmentation accuracy, the presence of brain shift resulting from the burr-hole craniotomy means that an interoperative data modality, such as test stimulation and/or electrophysiological recording,[26] is required to update the preoperative plan with an exact location of the STN boundary, assuming the electrode trajectory passes through the boundary.

Thus, the final positioning accuracy of the electrode depends on both the navigation (i.e., pre-operative segmentation) and targeting (i.e., intraoperative identification).[25,26]

4.5 *Future Directions*

In terms of future directions, automatic localization and segmentation methods have now reached a level in which subvoxel accuracy evaluation would be critical for further improving the state-of-the-art methods. Both the proposed method and other CNN-based approaches[6] (as well as the manual segmentations that these methods use for evaluation) consider the voxel to be the smallest unit, currently without the capability to determine subvoxel boundaries. Other, more invasive methods such as the use of implanted microelectode recordings (often used during DBS electrode implantations) could allow for finer spatial information to be collected,[27] but these would require a larger framework to register the spatial location of these electrodes into the MRI-space, especially considering brain shift.[3,28]

Another future direction would be to extend the localization and segmentation networks to multiple different subcortical anatomies to provide a more description patient-specific model for DBS electrode implantation planning. With the localization networks, this may even allow for the reversal of the registration-segmentation relationship: instead of using a registration algorithm to perform segmentation, the centroid and directional information for the various anatomies could be used as a series of corresponding points for simple point-based registration procedures such as thin-plate-splines.[29] This also raises some nuances in terms of HCI as participants may be more or less likely (depending on the structure) to correct the segmentation if a large number of structures are shown simultaneously, increasing the cognitive load.

Finally, we could in the future combine the feedback from the HCI experiment into the training itself, often called "human-in-the-loop" training. The benefit of this approach is that previously unannotated data could be used to further improve the network performance as users would be able to quickly annotate it via the correction mechanisms A and B in tandem. If the preannotated and unannotated datasets are tracked separately, there is a possibility that both an HCI evaluation as well as human-in-the-loop training could be performed simultaneously, although this may complicate the analysis of the former as the networks are not constant during the experiment.

5 Conclusions

This paper presents a two-part method for segmenting the left and right STNs from paired T1w and T2w MR images used for the pre-operative planning of DBS interventions. Each component is designed using CNNs, allowing it to learn from a relatively small number of available labeled datasets. The first part used a multiresolution CNN to determine an estimate of the STN location, which is then used to heavily crop the input images to the much smaller region of interest. The second network then directly segments these images using a U-Net style architecture. The benefit of this separation is twofold: it allows for more efficient use of computational resources, specifically memory allowing it to avoid patch processing and reconstruction;[6] and it provides an intermediate representation that a clinical user can easily correct, if needed. The localization network has an error that is consistently less than 6 mm, meaning that it provides a good initialization for the second algorithm that takes image patches of size $24 \times 24 \times 24$ mm, which can thus be understood to contain the entire STN. The final accuracy of the STN segmentation is $61.9 \pm 9.5\%$ Dice, which is almost equal to previously measured inter-rater variabilities,[20] a very high accuracy considering that the translation of the gold standard by a single pixel in the LR, AP, and SI directions results in Dice coefficients of $54.2 \pm 8.1\%$, $53.0 \pm 6.8\%$, and $55.0 \pm 5.5\%$, respectively.

This two-part algorithm also allows for additional interaction capabilities for clinicians through relocalization. This additional interaction is particularly important as it gives clinicians more autonomy and control over the segmentation problem while at the same time allowing for full automaticity. Our HCI experiment on novice users shows that the new interaction mechanism is highly usable and robust. This focus on designing machine learning methods with user interactivity in mind gives an example of how methods in medical image segmentation,

especially for small structures, should be implemented in the future, showing that the tradeoff between automaticity and interactivity is not a necessary one, but that even machine learning algorithms can be designed to be adaptable in real time.

Disclosures

The authors have no conflicts of interest to declare.

Acknowledgments

John S.H. Baxter was supported by the Insitut des Neurosciences Cliniques de Rennes (INCR). The authors would like to thank E. Maguet for his assistance in the earlier stages of this project's development as well as C. Haegelen for her assistance in data annotation and M. Corniola for his feedback on our HCI study interface design.

Data Statement

All data have been collected with informed patient consent. The data collection has been approved by the institutional ethics review board of the Centre Hospitalier Universitaire de Rennes.

References

1. M. S. Okun, "Deep-brain stimulation for Parkinson's disease," *N. Engl. J. Med.* **367**(16), 1529–1538 (2012).
2. T. D'Albis et al., "PYDBS: an automated image processing workflow for deep brain stimulation surgery," *Int. J. Comput. Assist. Radiol. Surg.* **10**(2), 117–128 (2015).
3. A. Horn and A. A. Kühn, "Lead-DBS: a toolbox for deep brain stimulation electrode localizations and visualizations," *Neuroimage* **107**, 127–135 (2015).
4. C. Haegelen et al., "Automated segmentation of basal ganglia and deep brain structures in MRI of Parkinson's disease," *Int. J. Comput. Assist. Radiol. Surg.* **8**(1), 99–110 (2013).
5. J. Bernal et al., "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," *Artif. Intell. Med.* **95**, 64–81 (2019).
6. F. Milletari et al., "Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound," *Comput. Vis. Image Underst.* **164**, 92–102 (2017).
7. M. Milchenko et al., "7t MRI subthalamic nucleus atlas for use with 3t MRI," *J. Med. Imaging* **5**(1), 015002 (2018).
8. J. Kim et al., "Automatic localization of the subthalamic nucleus on patient-specific clinical MRI by incorporating 7 t MRI and machine learning: application in deep brain stimulation," *Hum. Brain Mapp.* **40**(2), 679–698 (2019).
9. Y. Duchin et al., "Patient-specific anatomical model for deep brain stimulation based on 7 tesla MRI," *PLoS One* **13**(8), e0201469 (2018).
10. J. C. Lau et al., "Quantification of local geometric distortion in structural magnetic resonance images: application to ultra-high fields," *Neuroimage* **168**, 141–151 (2018).
11. W. Zhao et al., "Automated segmentation of midbrain structures in high-resolution susceptibility maps based on convolutional neural network and transfer learning," *Front. Neurosci.* **16**, 801618 (2022).
12. B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ants)," *Insight J.* **2**(365), 1–35 (2009).
13. D. Geffroy et al., "BrainVISA: a complete software platform for neuroimaging," in *Python in Neurosci. Workshop*, Paris (2011).
14. J. S. H. Baxter, E. Maguet, and P. Jannin, "Localisation of the subthalamic nucleus in MRI via convolutional neural networks for deep brain stimulation planning," *Proc. SPIE* **11315**, 113150M (2020).

15. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
16. E. Gibson et al., "Automatic multi-organ segmentation on abdominal CT with dense v-networks," *IEEE Trans. Med. Imaging* **37**(8), 1822–1834 (2018).
17. J. S. H. Baxter, E. Maguet, and P. Jannin, "Segmentation of the subthalamic nucleus in MRI via convolutional neural networks for deep brain stimulation planning," *Proc. SPIE* **11598**, 115981K (2021).
18. J. S. H. Baxter et al., "The semiotics of medical image segmentation," *Med. Image Anal.* **44**, 54–71 (2018).
19. S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): results of empirical and theoretical research," *Adv. Psychol.* **52**, 139–183 (1988).
20. W. H. Polanski et al., "Comparison of automatic segmentation algorithms for the subthalamic nucleus," *Stereotactic Funct. Neurosurg.* **98**(4), 256–262 (2020).
21. J. Kim et al., "Clinical deep brain stimulation region prediction using regression forests from high-field MRI," in *IEEE Int. Conf. Image Process. (ICIP)*, IEEE, pp. 2480–2484 (2015).
22. P. Nakkiran et al., "Deep double descent: where bigger models and more data hurt," arXiv:1912.02292 (2019).
23. B. R. Plantinga et al., "Individualized parcellation of the subthalamic nucleus in patients with Parkinson's disease with 7T MRI," *Neuroimage* **168**, 403–411 (2018).
24. J. Kim et al., "Semiautomatic segmentation of brain subcortical structures from high-field mri," *IEEE J. Biomed. Health Inf.* **18**(5), 1678–1695 (2014).
25. Y. Xiao et al., "Image guidance in deep brain stimulation surgery to treat Parkinson's disease: a comprehensive review," *IEEE Trans. Biomed. Eng.* **68**(3), 1024–1033 (2021).
26. A. E. Lang and H. Widner, "Deep brain stimulation for Parkinson's disease: patient selection and evaluation," *Movement Disorders* **17**(S3), S94–S101 (2002).
27. M. Peralta et al., "Sepaconvnet for localizing the subthalamic nucleus using one second micro-electrode recordings," in *42nd Annu. Int. Conf. IEEE Eng. in Med. and Biol. Soc. in Conjunction with the 43rd, Annu. Conf. Can. Med. and Biol. Eng. Soc.* (2020).
28. Y. Miyagi, F. Shima, and T. Sasaki, "Brain shift: an error factor during implantation of deep brain stimulation electrodes," *J. Neurosurg.* **107**(5), 989–997 (2007).
29. K. Rohr et al., "Landmark-based elastic registration using approximating thin-plate splines," *IEEE Trans. Med. Imaging* **20**(6), 526–534 (2001).

John S. H. Baxter is an INSERM researcher in the Laboratoire de Traitement du Signal et de l'Image (LTSI – INSERM UMR 1099) at the University of Rennes, France. Prior to that, he completed his PhD in medical imaging from the Western University, Canada. He is now part of the MediCIS group, where he researches the intersection of human–computer interaction and machine learning in medical image computing and computer-assisted interventions.

Pierre Jannin is an INSERM research director at the Medical School of the University of Rennes (France) and the head of the MediCIS Research Group in the LTSI (INSERM UMR 1099). He has more than 30 years of experience in designing and developing computer-assisted surgery systems. He was the president of the International Society of Computer-Aided Surgery and a board member of the MICCAI Society from 2014 to 2018. He is currently a SPIE senior member.